International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622
International Conference on Industrial Automation and Computing (ICIAC- 12-13th April 2014)

RESEARCH ARTICLE OPEN ACCESS

Concept Based Video Retrieval

Dipali N. Junankar *, Nita S. Patil **
*(Department of Computer Science, Datta Meghe College of Engineering Airoli Navi Mumbai
** (Department of Computer Science, Datta Meghe College of Engineering Airoli Navi Mumbai

ABSTRACT
Here common framework of concept-based video retrieval and several methods to improve the performance of the system are proposed. Features in color domain is calculated and utilized for detecting the key-frames and estimating the similarity between shots. By applying the predefined high-level rules, similar shots are merged and the scene boundaries are determined. The key frame is a simple and effective form of summarizing a long video sequence. Key frame extraction methods are used to obtain a set of frames that can efficiently represent and summarize video contents. An effective set of key frames are viewed as a high-quality summary of the video, should include the major objects and events of the video, and contain little redundancy and overlapped content. A region thesaurus that contains all the high-level features is constructed using a subtractive clustering method where each feature results as the centroid of a cluster. Then, a model vector that contains the distances from each region type is formed and a SVM detector is trained for each semantic concept.
Keywords **-** Feature Extraction, Keyframe extraction, Region Thesaurus

I. INTRODUCTION

The purpose for which a video is created is either entertainment, information, communication, or data analysis. For all these purposes, the user needs and demands vary substantially.

A consumer who wants to be entertained, for example, will be satisfied if a complete movie is accessible from an archive through a mobile phone. In contrast, a cultural anthropologist studying fashion trends of the eighties, a lawyer evaluating copyright infringement, or an athlete assessing her performance during training sessions might be more interested in retrieving specific video segments, without going through an entire video collection. For accessing complete video documents, reasonable effective commercial applications exist, YouTube and Netflix being good examples. Video search applications for consumers and professionals targeting at retrieval of specific segments, however, are still in a nascent stage . Users requiring access to video segments are hardly served by present-day video retrieval applications.

We review video search solutions that target at retrieval of specific segments. Since humans

II. REVIEW OF LITERATURE

When video is brought for processing first the boundary of every shots in video is detected using color histogram, using edge detection, Using segmentation.

In color histogram shot boundary detection **e**ach (352 * 288) frame of the video and computes a color histogram using a total of 192 different colour bins where each bin contains the percentage of pixels from the whole frame. When this is done, the color histogram for each frame is checked against the histogram for the one following and a similarity metric is used to compute the likeness between the two adjacent frames. When the sequence of similarity values is analysed, a shot will show a big change in the sequence where the shot boundary occurs, so by looking for these peaks in the differences, a shot boundary can be detected.

Edge detection approach looks not at the colour differences, but at the differences between the edges detected in adjacent frames. Each frame in the digital video is turned into a greyscale image and Sobel filtering applied to detect edges. The method looks for similar edges in adjacent frames to detect a shot boundary. The principle behind the edge detection approach is that it can counter problems caused by fades and dissolves and other transitions which are invariant to gradual colour changes. With edge detection, even when there are gradual transitions between shots, there should always be a pair of adjacent frames where an edge is detected in one, and not in the other and identifying an occurrence of this on a large scale locates a shot transition. Like the colour-based method above, this will require a minimum difference between adjacent frames to detect a shot cut but it has the advantage of not being fooled by a large colour change.

With the detected shots from the whole film, a high-level feature extraction for structuring the video is needed to collect the most characterized

information rather than using all video frames to avoid the costly computation and speed up the system. Key frame extraction can be done using dominant set clustering, using shot based key frame extraction, using cluster based extraction, using genetic algorithm.

The concept of dominant set provides an effective framework for iterative pairwise clustering. Considering a set of samples, an undirected edge-weighted graph with no self-loops is built in which each vertex represents a sample and two vertices are linked by an edge whose weight represents the similarity of the two vertices. To cluster the samples into coherent groups, a dominant set of the weighted graph is iteratively found and then removed from the graph until the graph is empty. Different from traditional clustering algorithms, the dominant-set clustering automatically determines the number of the clusters and has low computational cost.

In this algorithm video is first segmented into shots and then key frames are extracted for each shot independently. Key frame extraction techniques can be roughly categorized into sequential and cluster-based methods. In sequential methods, consecutive frames are compared in a sequential way and key frames are detected depending on the similarity with either the previous frames or the previously detected key frame. Although the sequential methods consider the temporal ordering among frames, they only compute the similarity between adjacent frames and ignore the overall change trend in the shot range in cluster-based methods, the frames are grouped into a finite set of clusters in the selected feature space, and then the key frame set is obtained by collecting the representatives of each cluster. In this method, temporal information of the frames is not considered; that is, key frames are selected regardless of the temporal order of each frame. If key frames are extracted for each shot independently and the scenery changes slowly in each shot, cluster-based methods are able to provide an understanding of the overall visual content of a video. In this way, the number of key frames in each shot is compact, capturing adequately the content variation along a video

GA (Genetic Algorithm) is modeled on the mechanism of biological genetics and natural selection. It is a kind of optimal search algorithm by artificial method. It is a mathematical simulation of biological evolution, and is one of the most important forms of evolutionary computation.

GA is mainly composed by encode and decode, fitness function and genetic manipulation. Coding is to express the chromosome of the individuals in biotic population as binary sequence, while decoding is the reverse process of coding.

Fitness function is a function which is used to measure the individuals' degree of merits. The fitness function $F(k)=k*d(1,k)+(L-k)*d(k, L)$ is used as Genetic manipulation contains selection crossover mutation.

Feature extraction is done on the extracted key frames. There are different methods for feature extraction like RGB moments, HSV color correlogram.

In RGB Moment the frame is divided into 5x5 blocks and the first, second, and third central moments of RGB components are calculated for each block. And then all the moments are concatenated in order of blocks to form a feature with 255 dimensions

In RGB_BlkHist each component of RGB color space is quantized to 2 bits, and a histogram with 64 bins is generated. To enhance the invariance to space location, the frame is divided into 3x3 blocks, and the histogram is calculated on each block. At last, all the histograms are cascaded into one, which is of 3x3x64=576 dimensions.

In HSV_CorHist: a histogram of color correlograms in a frame. In HSV color space, the H, S and V components are quantized to 16, 4 and 4 bits respectively and then two HSV correlograms are computed at the distances 1 and 3. Finally, a feature vector with 512 dimensions is made by joining the two correlograms together.

Gabor & Gabor Sort is a common texture feature usually used in face recognition , and an improvement is made in our system to fit the video retrieval. The feature is calculated in 3 scales and 6 orientations, and the statistical characters (e.g. mean and variance) are computed on 3x3 blocks which is split from one frame, and then a feature vector with 324 dimensions is formed. The Gabor_Sort, similar with Gabor feature, is a sorted feature according to the means of blocks[11].

In EDH (Edge Directional Histogram) the edge histogram represents the distribution of five types of edges, namely four directional edges (0°, 45°, 90°, 135°) and one nondirectional edge, which is recommended in MPEG-7. In our framework, the frame is divided into 392 3x4 blocks, and then EDHs of the 12 blocks are cascaded into a 60 dimensional feature vector. HOG_Blk & HOG_Edge: HOG is an effective but not efficient feature, which calculates the histogram of oriented gradients and is often used in human detection To reduce the complexity, we use HOG on nonoverlapped blocks. For HOG_Edge, we calculate it on edges of objects and adapt the BoW method to handle the unfixed points, which is similar to the SIFT feature in a sense. LBP is a local texture feature which is proposed by Ojala in 1996. However, there is some difference with Ojala's work in our system where the feature is described on the edge of an image and a 256-D histogram is generated

to represent the image[13].

III. PROPOSED SYSTEM

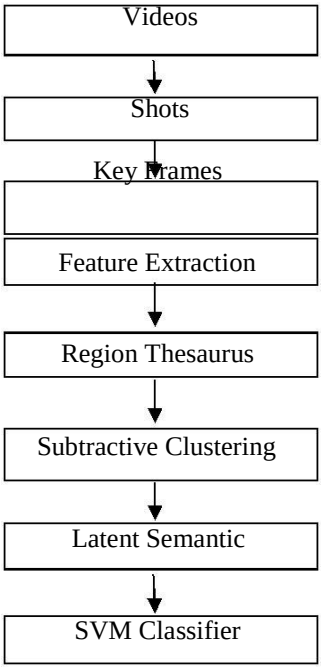

Fig1. Propose System

A. Video Shot Boundary Detection Using Color Histogram

The approach here takes each (352 * 288) frame of the video and computes a color histogram using a total of 192 different color bins where each bin contains the percentage of pixels from the whole frame. When this is done, the color histogram for each frame is checked against the histogram for the one following and a similarity metric is used to compute the likeness between the two adjacent frames. When the sequence of similarity values is analyzed, a shot will show a big change in the sequence where the shot boundary occurs, so by looking for these peaks in the differences, a shot boundary can be detected. Euclidean distance measure is used to find the histogram difference [3].

B. Cluster Based Extraction for Key Frame Extraction

In cluster-based methods, the frames are grouped into a finite set of clusters in the selected feature space, and then the key frame set is obtained by collecting the representatives of each cluster. In this method, temporal information of the frames is not considered; that is, key frames are selected regardless of the temporal order of each frame. If key frames are extracted for each shot independently and the scenery changes slowly in each shot, cluster-based methods are able to provide an understanding of the overall visual content of a video. In this way, the number of key frames in each shot is compact, capturing adequately the content variation along a video [4].

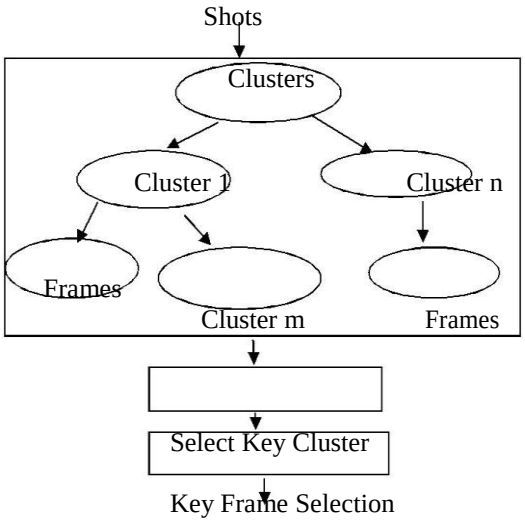

Fig.2 Key Frame Extraction using Clustering

C. Feature Extraction

Color Extraction

The standardized MPEG-7 Dominant Color Descriptor is formed after the clustering of the present colors within an image or a region of interest. This way, the representative colors of each keyframe are calculated. The selected low-level visual features of the image consist of the representative (dominant) colors, their percentages in the region, and optionally their spatial coherencies and their variances. What discriminates the use of the dominant color description of an image, instead of i.e. a color histogram is that the representative colors are computed each time, based on the features of the given image rather than being fixed in the color space [5].

The color description is then formed as follows: DCDN = [fC1; P1g; fC2; P2g; : : : ; fCN; PNg]

RGB Moment

The frame is divided into 5x5 blocks and the first, second, and third central moments of RGB components are calculated for each block. And then all the moments are concatenated in order of blocks to form a feature with 255 dimensions

Feature Extraction

For extracting edge features for an image, it

is divided into sub-images, called as blocks. A block is characterized by generating a histogram of its edge distribution. There are 3 kinds of block; global, semi-global and local block. The histograms for each block could also represent the occurrence frequencies for 5 types of edge, called as bin: vertical, horizontal, 45-degree diagonal, 135-degree diagonal and non-directional edge bin. The local blocks are generated by dividing an image into 4×4 non-overlapping blocks. We can also produce 13 semiglobal blocks by combining 4 successive local blocks. Therefore, an image is represented by 30 blocks (1 global block, 13 semi-global blocks and 16 local blocks) containing 150 bins (30 blocks × 5 bins/block). The global and semi-global bock bins may considerably reflect edge distributions for the whole image [6].

The texture description for a region of an image is then formed as follows:

EDH = [fDC; fSD; e1; e2; :::; e30; d1; d2; :::; d30]

D. Region Thesaurus

Generally, a thesaurus combines a list of every term in a given domain of knowledge and a set of related terms for each term in the list. Here the constructed Region Thesaurus contains all the Region Types that are encountered in the training set. These region types are the centroids of the clusters and all the other feature vectors of a cluster are their synonyms. It is important to mention that when two region types are considered to be synonyms, they belong to same cluster, thus share similar visual features, but do not necessarily share the same semantics. By using a significantly large training set of keyframes, our thesaurus is constructed and enriched [5].

E. Subtractive Algorithm

Suppose we don't have a clear idea how many clusters there should be for a given set of data. *Subtractive clustering*, is a fast, one-pass algorithm for estimating the number of clusters and the cluster centers in a set of data. The cluster estimates obtained from the subclust function can be used to initialize iterative optimization-based clustering methods and model identification methods. The subclust function finds the clusters by using the subtractive clustering method. This algorithm can have a large reduction on the number of training samples, based on the density of surrounding data points. Namely, all data points in a small dense zone of one point center will be replaced by this typical one. On the other hand, the sparse points in the input space will remain as cluster centers themselves. So this algorithm is noise robust, outliers have little influence on the choice of cluster centers [7].

F. Latent Semantic Analysis

Latent Semantic Analysis (LSA) is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text. The underlying idea is that the aggregate of all the word contexts in which a given word does and does not appear provides a set of mutual constraints that largely determines the similarity of meaning of words and sets of words to each other. The adequacy of LSA's reflection of human knowledge has been established in a variety of ways. For example, its scores overlap those of humans on standard vocabulary and subject matter tests; it mimics human word sorting and category judgments; it simulates word–word and passage– word lexical priming data; and, it accurately estimates passage coherence, learnability of passages by individual students, and the quality and quantity of knowledge contained in an essay. After processing a large sample of machine-readable language [8].

Latent Semantic Analysis (LSA) represents the words used in it, and any set of these words such as a sentence, paragraph, or essay either taken from the original corpus or new, as points in a very high (e.g. 50-1,500) dimensional "semantic space".

G. Support Vector Machine

Support vector machines (SVMs) are a kind of machine learning method for both classification and regression problems. They base on the structural risk minimization principle. SVMs have been applied to deal with a wide range of problems due to their high generalization ability and good classification precision. At present, SVMs have been applied to many classification and recognition fields, such as handwriting recognition, text classification, face recognition, and speech recognition, etc.

Support vector machines outperform conventional classifiers especially when the number of training data is small. However, for the large and high dimensional data sets, the kernel computation and optimization time for training a SVM are time consuming [9].

IV. CONCLUSION

When we use content base video retrieval sometimes we get irrelevant results. To remove the drawback of CBVR we used concept base video retrieval. The video is divided into shots using the histogram of the scene. Shots are further divided into frames using dominant set clustering. To find the color features of the keyframe we use RGB Moment

method. Shape of the keyfrme is found out using Edge Directional Histogram. The region thesaurus is used to store the keyframes under different concepts and the videos are retrieve. Thus using concept based video retrieval we try to minimize the semantic gap between low level features and the high level features.

REFERENCES

[1]. Concept-Based Video Retrieval By Cees G. M. Snoek and Marcel Worring
[2]. An Improved System For Concept-Based Video Retrieval Xin Guo, Zhicheng Zhao, Yuanbo Chen, Anni Cai
[3]. Review on different methods of video shot boundary detection,Ravi Mishra & Sanjay Kumar Singhai, International Journal of Electrical and Electronics Engineering IJEEE Vol.1, Issue 1 Aug 2012 46-57
[4]. Key-Frame Extraction Using Dominant-Set Clustring, Xianglin Zeng, WeimingHu, Wanqing Liy, Xiaoqin Zhang, Bo Xu, NSFC (Grant No. 60672040
[5]. High-Level Concept Detection in Video Using a Region Thesaurus, Evaggelos Spyrou and Yannis Avrithis
[6]. Design and Implementation of a Concept-based Image Retrieval System with Edge Description Templates**,** J. H. Choi, S. H. Park a, S. J. Park, SPIE Vol. 5307 © 2004 SPIE and IS&T · 0277-786X/04/$15
[7]. Classifier Fusion for SVM-Based Multimedia Semantic Indexing, St´ephane Ayache, Georges Qu´enot, J´erˆome Gensel
[8]. An Introduction to Latent Semantic Analysis, Landauer T. K., Foltz, P. W., & Laham, D., 1998
[9]. Domain Transfer SVM for Video Concept Detection, Lixin Duan Ivor W. Tsang Dong Xu, Stephen J. Maybank, 978-1-4244-3991-1/09/$25.00 ©2009 IEEE